GN Docket No. 14-28

BITAG

Broadband Internet Technical Advisory Group

Real-time Network Management of Internet Congestion

A BROADBAND INTERNET TECHNICAL ADVISORY GROUP
TECHNICAL WORKING GROUP REPORT

A Uniform Agreement Report

Issued:
October 2013

About the BITAG

The Broadband Internet Technical Advisory Group (BITAG) is a non-profit, multi-stakeholder organization focused on bringing together engineers and technologists in a Technical Working Group (TWG) to develop consensus on broadband network management practices and other related technical issues that can affect users' Internet experience, including the impact to and from applications, content and devices that utilize the Internet.

The BITAG's mission includes: (a) educating policymakers on such technical issues; (b) addressing specific technical matters in an effort to minimize related policy disputes; and (c) serving as a sounding board for new ideas and network management practices. Specific TWG functions also may include: (i) identifying "best practices" by broadband providers and other entities; (ii) interpreting and applying "safe harbor" practices; (iii) otherwise providing technical guidance to industry and to the public; and/or (iv) issuing advisory opinions on the technical issues germane to the TWG's mission that may underlie disputes concerning broadband network management practices.

The BITAG Technical Working Group and its individual Committees make decisions through a consensus process, with the corresponding levels of agreement represented on the cover of each report. Each TWG Representative works towards achieving consensus around recommendations their respective organizations support, although even at the highest level of agreement, BITAG consensus does not require that all TWG member organizations agree with each and every sentence of a document. The Chair of each TWG Committee determines if consensus has been reached. In the case there is disagreement within a Committee as to whether there is consensus, BITAG has a voting process with which various levels of agreement may be more formally achieved and indicated. For more information please see the BITAG Technical Working Group Manual, available on the BITAG website at www.bitag.org.

BITAG TWG reports focus primarily on technical issues. While the reports may touch on a broad range of questions associated with a particular network management practice, the reports are not intended to address or analyze in a comprehensive fashion the economic, legal, regulatory or public policy issues that the practice may raise.

BITAG welcomes public comment. Please feel free to submit comments in writing via email at comments@bitag.org.

Executive Summary

The Internet, as is the case with many other networks such as highways and electricity grids, operates under the assumption that capacity will be set to a level such that total peak demand will occasionally exceed capacity. Further, the Internet is designed so that multiple users may dynamically share capacity and multiple services may share the same network links and routers, which is more efficient than offering individual users dedicated capacity or different services using separate links and routers.

Every link and router in the various networks that make up the Internet has a limit on its capacity to handle data. The capacity of each link and router in individual networks is determined by the equipment installed by the entity that runs each network in an attempt to optimize performance and cost; the lower the capacity relative to expected demand, the greater the probability that demand upon that link or router at times may exceed its capacity.

Significantly, a user's instantaneous demand for broadband Internet is bursty, meaning that it changes rapidly in time – and when aggregate instantaneous demand exceeds capacity on a network it causes congestion, which can degrade performance.

Network operators typically estimate demand months to years in advance, and use such demand estimates to plan a schedule for capacity upgrades. Since it may take months to implement a capacity upgrade, the time scale for managing congestion in this manner is months to years. Thus, although capacity planning can greatly affect how much congestion occurs on a network over time, it cannot react to congestion as it occurs.

The impact of congestion upon applications depends on the duration of congestion – which can vary from thousandths of a second up to hours or more – and the nature and design of the application. If the duration of congestion is short enough or the application is tolerant enough of congestion, a user will not notice any degradation in performance. Congestion is thus a problem only when its duration is long enough to be disruptive to applications. Congestion in a network can occur for a wide variety of reasons, some of which can be anticipated and some of which cannot.

This report describes how network resources are allocated on a short time scale in order to, among other objectives, manage congestion on the network, and how such congestion management impacts applications and users.

Congestion management practices are an important subset of network management practices implemented by a variety of parties or organizations, including Internet Service Providers (ISPs) and Application Service Providers (ASPs). Policymakers have expressed great interest in learning what congestion management practices are used in the Internet and how these practices impact users and the broader Internet ecosystem. Furthermore, an understanding of congestion management techniques and practices is crucial in discussions about reasonable network management.

One of the key design questions about any congestion management practice concerns the subset of network traffic to which the practice is applied, and its impact upon users and applications. Network operators apply some practices to all traffic on their networks, whereas in other cases practices are applied only to the traffic of specific users or to the traffic associated with specific applications. Application- or user- based congestion management practices may achieve better performance for selected applications. They also may enable service providers to offer connectivity products that cater to particular customer's tastes or needs. However, they add complexity, which may result in added costs that each network operator will evaluate. In some cases application- or user- based congestion management practices may be harmful to applications.

Congestion management practices are composed of generic technical building blocks, described in this report as traffic management "techniques". This report discusses a range of user- and application- based congestion management techniques, including classification of packets, reservation of resources for particular network flows, storage of content in multiple locations, rate control, routing and traffic engineering, packet dropping, and packet scheduling.

Congestion management techniques may be combined to offer a collection of capabilities in various network architectures, and can create services with differentiated performance either within an operator's network or end-to-end. There are also architecture-specific implementations of congestion management techniques for broadband Internet access over cable, telephone, and cellular networks and for Content Delivery Networks. The offerings of a service provider often include multiple services that may utilize the same network links and routers. While there are benefits and efficiencies to sharing capacity between multiple services, such sharing of capacity also requires the use of congestion management practices.

Congestion management "practices" are the uses of particular techniques by particular network operators to avoid, limit, or manage congestion. This report illustrates a range of congestion management practices that show how providers may combine user- or application- based congestion management techniques, including traffic shaping, prioritization, transcoding, resource reservation, and preferential treatment.

The report begins in Sections 1 and 2 by giving an overview of congestion and BITAG's interest in the issue. Section 3 defines congestion and describes instances in which congestion can occur, the locations in the network where congestion can occur, the indicators of congestion, and the impact congestion can have on applications.

In Section 4, the report articulates the differences between congestion management techniques and congestion management practices, and describes the different time scales at which congestion can be seen to occur in the network. This section also describes the parties that implement congestion management practices and on what basis.

Although all congestion management is important, in order to limit scope and length Sections 5-7 focus on congestion management techniques and practices that: (1) are

implemented or potentially implemented in a network that supports consumer broadband Internet access services; (2) act on a time scale of minutes or less; (3) are used for purposes of congestion management; and (4) are based on user or application.

In Section 5, the report focuses on specific congestion management techniques. Section 6 gives specific examples of congestion management practices that are based on user or application. Finally, Section 7 gives the Technical Working Group's recommendations.

At a high level, the recommendations of BITAG's Technical Working Group are:

- **ISPs and ASPs should disclose information about their user- or application-based network management and congestion management practices for Internet services in a manner that is readily accessible to the general public.** This information should be made available on network operators' public web sites and through other typically used communications and channels, including mobile apps, contract language, or email. ISPs and ASPs may choose to use a layered notice approach, using a simple, concise disclosure that includes key details of interest to consumers complemented by a more thorough and detailed disclosure for use by more sophisticated users, application developers, and other interested parties. The detailed disclosure should include: descriptions of the practices; the purposes served by the practices; the types of traffic subject to the practices; the practices' likely effects on end users' experiences; the triggers that activate the use of the practices; the approximate times at which the practices are used; and which subset of users may be affected. The disclosures should also include the predictable impact, if any, of a user's other subscribed network services on the performance and capacity of that user's broadband Internet access services during times of congestion, where applicable.

- **Network operators should use accepted industry "Best Practices," standardized practices, or seek industry review of practices.** Network standards setting organizations and technical industry bodies produce considered recommendations of Best Practices and standard practices for a variety of operational issues including congestion and congestion management. Where network operators see the need for an innovative solution that has not been standardized or documented as a Best Practice, these network operators should consider bringing their unique network or congestion management practices to such groups for discussion and documentation.

- **When engaging in a congestion management practice that could have a detrimental impact on the traffic of certain users or certain applications, the practice should be designed to minimize that impact.** Some congestion management practices may cause certain users or certain applications to experience performance degradation. ISPs and ASPs should seek to minimize such degradation to the extent possible while still managing the effects of the congestion that originally triggered the use of the practice.

- **If application-based congestion management practices are used, those based on a user's expressed preferences are preferred over those that are not.** User- and application- agnostic congestion management practices are useful in a wide variety of situations, and may be sufficient to accommodate the congestion management needs of network operators in the majority of situations. However, at times network operators may choose to use application-based congestion management practices, in which case those that prioritize application traffic according to a user's expressed preferences are preferred over those that do not.

- **If application-based criteria are used by a network operator, they should be tested prior to deployment and on an ongoing basis.** Application-based classification by network operators (e.g., using deep packet inspection) can sometimes be erroneous. If network operators choose to use application-based criteria for congestion management, the accuracy of the classifier should be tested before deployment.

- **ASPs and CDNs should implement efficient and adaptive network resource management practices.** ASPs and CDNs should match use of network resources to the performance requirements of the application. Applications should be designed to efficiently and adaptively use network resources, to the extent feasible given the application's requirements.

Table of Contents

1. Issue Overview

The Internet, as is the case with many other networks such as highways and electricity grids, operates under the assumption that capacity will be set to a level such that total peak *demand* will occasionally exceed *capacity*. Further, the Internet is designed so that multiple users may dynamically share capacity and multiple services may share the same network links and routers, which is more efficient than offering individual users dedicated capacity or different services using separate links and routers.

Every link and router in the various networks that make up the Internet has a limit on its capacity to handle data. The capacity of each link and router in individual networks is determined by the equipment installed by the entity that runs each network in an attempt to optimize performance and cost; the lower the capacity relative to expected demand, the greater the probability that demand upon that link or router at times may exceed its capacity.

Significantly, a user's instantaneous demand for broadband Internet is *bursty*, meaning that it changes rapidly in time – and when aggregate instantaneous demand exceeds capacity on a network it causes *congestion*, which can degrade performance.

Network operators typically estimate demand months to years in advance, and use such demand estimates to plan a schedule for capacity upgrades. Since it may take months to implement a capacity upgrade, the time scale for managing congestion in this manner is months to years. Thus, although capacity planning can greatly affect how much congestion occurs on a network over time, it cannot react to congestion as it occurs.

The impact of congestion upon applications depends on the duration of congestion – which can vary from thousandths of a second up to hours or more – and the nature and design of the application. If the duration of congestion is short enough or the application is tolerant enough of congestion, a user will not notice any degradation in performance. Congestion is thus a problem only when its duration is long enough to be disruptive to applications. Congestion in a network can occur for a wide variety of reasons, some of which can be anticipated and some of which cannot.

This report describes how network resources are allocated on a short time scale in order to, among other objectives, manage congestion on the network, and how such congestion management impacts applications and users.

One of the key design questions about any congestion management practice concerns the subset of network traffic to which the practice is applied, and its impact upon users and applications. Network operators apply some practices to all traffic on their networks, whereas in other cases practices are applied only to the traffic of a subset of specific users, to a subset of types of applications, to all instances or specific instances of applications, or to specific components of such applications. Application- or user- based congestion

management practices may achieve better performance for various application-related traffic. Such practices also may enable service providers to offer connectivity products that cater to particular customer's tastes or needs. However, they add complexity, which may result in added costs that each network operator will evaluate. In some cases application- or user- based congestion management practices may be harmful to applications.

This report focuses on real-time Internet traffic management practices based on users or applications that are used on networks operated by Internet Service Providers (ISPs) and Application Service Providers (ASPs) (known as "network operators" throughout this report) for the purposes of congestion management.[1] Network management practices used by network operators for purposes other than congestion management are outside the scope of this report. Practices that are not implemented in real-time are also outside the scope, including usage caps and usage charges.

The analysis distinguishes between traffic management "techniques," which are generic technical building blocks, and traffic management "practices," which are the applications of particular techniques by particular network operators to avoid, limit, or manage congestion. With respect to techniques, the analysis considers where in the network, and at which layer, a traffic management technique is applied, and what type of traffic management functionality is applied. With respect to practices, the analysis considers who decides whether a traffic management practice is applied and on what basis. It is important to examine the criteria and indicators of congestion that trigger a practice.

2. BITAG Interest in the Issue

Congestion management practices are an important subset of network management practices implemented by a variety of parties or organizations, including Internet Service Providers (ISPs) and Application Service Providers (ASPs). Policymakers have expressed great interest in learning what congestion management practices are used in the Internet and how these practices impact users and the broader Internet ecosystem [FCC 07-31].

Policymakers often comment that network architectures and technologies may impact congestion management practices, but are looking for guidance as to how this occurs. Furthermore, an understanding of congestion management techniques and practices is crucial in discussions about reasonable network management.

---

[1] For purposes of this report, an Internet Service Provider (ISP) is defined as a provider of broadband Internet access service, an Application Service Provider (ASP) is defined as a provider of applications used on broadband Internet access services, and a network operator is defined as an ISP, or an ASP that operates a network. Some ASPs operate networks that interconnect with ISPs, while other ASPs attach servers directly to ISPs.

3. Characterization of Congestion

It is important to understand what is meant by the term "congestion" in the context of the Internet, and this section of the report provides an overview. Section 3.1 discusses demand, capacity, and congestion. Section 3.2 describes in what instances congestion can occur. Section 3.3 describes the locations in the network where congestion can occur. Section 3.4 describes the indicators of congestion. Section 3.5 describes the impact congestion can have on applications.

3.1. Definition of congestion

As pictured in Figure 1, a user's instantaneous *demand* for broadband Internet (measured in bits per second) is *bursty*, meaning that instantaneous demand changes rapidly in time.[2] A user's average demand, measured over several days, is much lower than the user's peak demand. The Internet is designed so that multiple users may dynamically share *capacity*, a concept called *statistical multiplexing*.

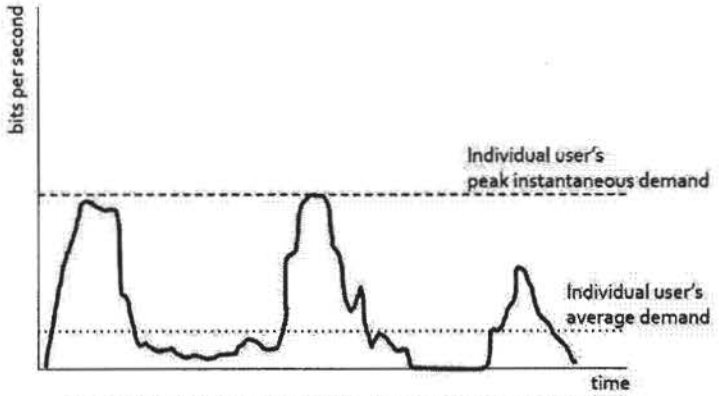

Figure 1. A single user's demand for broadband Internet, measured in bits per second.

Figure 2 illustrates two users sharing capacity. One user's instantaneous demand is shown as a solid black line, another user's instantaneous demand as a solid grey line, and the sum of their instantaneous demands as a dashed black line. The total average demand is simply the sum of the user's individual average demands. However, users' individual instantaneous demands are usually uncorrelated with each other, so that they burst to high levels at different times. As a result, the total peak demand (the highest point on the dashed curve) is usually far less than the sum of users' individual peak

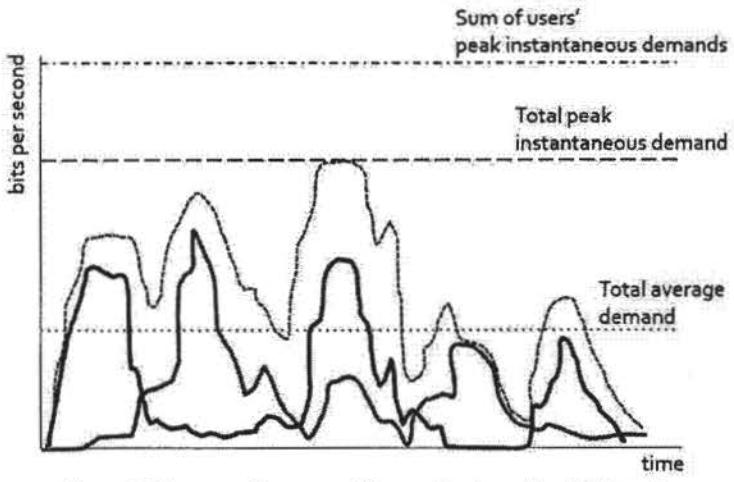

Figure 2. The sum of two users' demand for broadband Internet.

[2] Figures 1-3 are for illustrative purposes only, and do not represent actual measured network data.

demands (the dash dot line at the top of the figure) due to the fact that users' individual peaks are typically non-concurrent. Some of the reasons for occurrences of high demand and of fluctuation in demand, along with the potential duration of such demand, are described in Section 3.2.

All links and routers in a network have a limit on their capacity to handle data, as described in Section 3.3. As illustrated in Figure 3, for purposes of this report, *congestion* is defined as the effect upon network performance during time periods in which instantaneous demand exceeds capacity.[3]

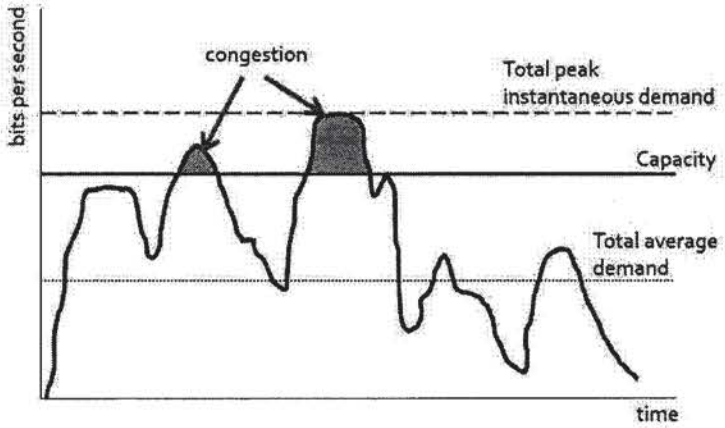

Figure 3. Congestion occurs when instantaneous demand exceeds capacity.

The Internet operates under the assumption that capacity will be set to a level such that total instantaneous peak demand will occasionally exceed capacity, see e.g. [Kurose and Ross, section 3.1]. This design is based upon the cost efficiency that can be gained through dynamic sharing of capacity. The capacity of each link and router in a network is determined by the equipment installed by the entity that runs the network in an attempt to optimize performance and cost; the lower the capacity relative to expected to demand, the greater the probability that instantaneous demand upon that link or router may at times exceed its capacity. Because not all users are active or fully use their maximum Internet connection speed at the same time, a network operator may install capacity in a link or router at a level above the total average demand but below the total instantaneous peak demand. This well-established practice of network design lowers the cost of creating a network and providing connectivity. It is used not only in the Internet, but also on highways, electricity networks, and air transportation networks, since it would be prohibitively expensive to add enough capacity to ensure that congestion never occurs.

Congestion may cause an increase in the *end-to-end packet delay*, which is the delay from the time a packet is transmitted by the source until it is received by the destination. Congestion may also cause an increase in *end-to-end packet loss,* which is the proportion of packets that do not arrive at the destination. These indicators of congestion and methods for network operators to measure congestion are discussed in Section 3.4.

---

[3] Alternate definitions of *congestion* include the effect upon network performance during time periods when (1) average demand exceeds capacity over a specified measurement interval, (2) the load over a specified measurement interval exceeds a specified threshold, and (3) packets are dropped by a router [Evolution of Internet Congestion]. These indicators of congestion are discussed in Section 3.4.

The duration of congestion can vary from milliseconds (thousandths of a second) to hours. The impact of congestion upon applications depends on the duration and severity of congestion and the nature and design of the application. If the duration of congestion is short enough or the application is tolerant enough of congestion, a user will not notice any degradation in performance. Congestion is thus a problem only when its duration is long enough to be disruptive to applications. The impact that congestion has on users, applications, and ASPs is discussed in Section 3.5.

The total average demand across an operator's network varies by hour and day of the week. For consumer wireline networks, average demand is usually highest during the evening hours of each day, typically exhibiting a pattern similar to that illustrated in Figure 4.

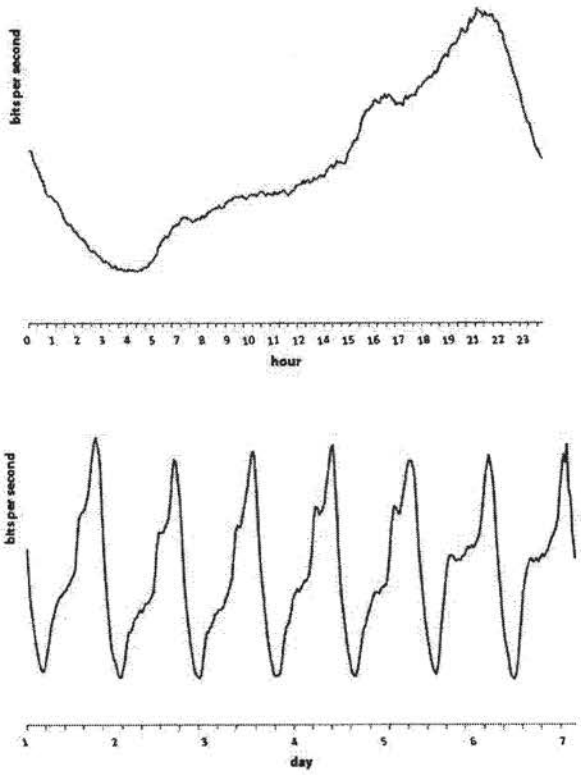

Figure 4. Total average demand by hour and day of week.

If capacity sufficiently exceeds the total average demand during the busiest hours, then the duration of congestion is generally short – milliseconds to seconds. This is the desired situation, as most users and applications will not experience a reduction in perceived performance. Occasional short-term congestion is unavoidable.

In contrast, if capacity does not sufficiently exceed the total average demand during the busiest hours, then the duration of congestion may be longer – minutes to hours – which will significantly degrade the perceived performance of most users and applications. In this situation, the only effective solutions to long-term congestion are to either increase capacity or decrease demand (which are not core topics discussed further in this document). A network operator will usually schedule upgrades to the links and routers to increase capacity months before it predicts that such long-term congestion will occur [RFC 6057]. The cost of adding capacity varies according to the technology, and is generally the highest in access networks (which include the portions of the network often referred to as the "last mile"). A network operator will consider cost and performance using each particular access technology when deciding how and when to increase capacity. Reductions in average demand can, among other ways, be accomplished by creating more bandwidth-efficient applications and services or altering users' incentives through pricing plans.

3.2. Occurrence and Duration of Congestion

Congestion in an ISP or ASP network can occur for a wide variety of reasons, some of which can be anticipated and some of which cannot. For purposes of this report, the causes of congestion can roughly be classified as: recurrent congestion, predictable events, unpredictable events, and random congestion.

3.2.1. Recurrent Congestion

The normal patterns of human and business activities create cyclical and recurring time periods when overall traffic on a network significantly increases. For example, parts of an ISP network with a high concentration of business users are likely to see higher usage during business hours than at other times. In contrast, parts of the network with a high concentration of residential users are likely to see higher usage during evening hours, as illustrated previously in Figure 4.

Recurrent congestion will typically last for multiple hours and when it occurs generally displays a periodic pattern, for example, weekday afternoons, every evening, or some other recurrence clearly linked to underlying human behaviors. As a result, this type of congestion tends to be predictable. Comparing Internet data networks to highways, an analogy for recurrent congestion is the average traffic patterns according to the time of day, including normal delays during rush hours.

3.2.2. Predictable Events

Specific predictable events can be the cause of network congestion by creating unusual Internet demand in addition to the existing demand, either as sources or destinations of traffic. Again comparing Internet data networks to highways, an analogy for congestion caused by predictable events is the incremental traffic caused by planned events such as road construction. A variety of examples can illustrate this type of congestion in the Internet:

- ***Mass in-person event***

  A mass in-person event occurs where many people gather in one physical place: sporting events, conventions, or political rallies, for example. These events are more likely to occur in populated areas. Wireless networks are most prone to these events, since they typically involve network users physically coming together. These events generally last for hours, although some may last only a few minutes. Some events are planned far in advance, and are thus predictable, but some occur with little warning.

- ***Mass on-line event: Users accessing an Internet site***

  A mass online event occurs when many people try to reach a particular Internet destination at the same time or try to consume the same streamed content from a single source: streamed sporting events, live news events, the release of popular software,

online ticket sales of popular events, or shopping at popular websites on major holidays, for example. This type of event is the virtual version of a mass in-person event. While the traffic to or from any individual user may be relatively small, the concentration of traffic at the destination may be large and thereby cause congestion. These events generally last for hours to days. While many mass on-line events are predictable, the popularity of any given event can be unpredictable.

- ***Mass on-line event: User-to-user communication***

A mass distributed online event occurs when many users try to communicate directly with each other at the same time, for example during major holidays. This type of event can cause congestion over a wide geographic region. Generally these events last for hours and are predictable.

3.2.3. Unpredictable Events

Specific unpredictable events can also be the cause of network congestion. A highway analogy for congestion caused by unpredictable events is the incremental traffic caused by unplanned events such as accidents. A variety of examples can illustrate this type of congestion in the Internet:

- ***Changes in routing***

While changes in routing of traffic typically decrease congestion, unexpected changes in routing of high volumes of traffic can increase congestion, e.g. when a large content provider changes the ISP from whom it purchases Internet access. These events may cause congestion at the boundary between two ISPs' networks, as the increased flow may exceed capacity. Addressing such congestion may involve manual changes to routing (on a time scale of hours) or discussions about interconnection agreements between the two ISPs (on a time scale of days to months).

- ***Emergencies***

Unexpected life-threatening or property-damaging events – earthquakes, hurricanes, floods, tornadoes, or major automobile traffic incidents, for example – can cause large increases in network traffic. The traffic comes both from the direct response to the emergency and often from the desire of users not directly impacted by the emergency to seek information. Dramatic weather can also shift demand among networks, for example because of people working from home due to impassable roads. These events tend to be localized, although the size of the geographic area covered can vary greatly depending on the nature of the emergency. While many such events last for hours, some can persist for days or weeks.

- ***Network Accidents and Failures***

Congestion can occur because of a temporary loss of capacity in the network due to failures or accidents. As with any technology, links or routers can fail. Failures may be

due to hardware, the result of software bugs, or secondary effects of an emergency that causes a loss of power or spikes in power that damage equipment. For example, network links can be severed due to earthquakes, high winds, tornadoes, floods, fallen trees, construction activities, and automobile accidents.

The loss of network links or routers results in a decrease in capacity. The remaining parts of the impaired network may not have sufficient capacity to accommodate the redirected traffic, and congestion may result. In addition, portions of the network may lose connectivity, which can create further congestion if sources retransmit packets that did not arrive at their destinations.

Congestion resulting from accidents and failures can last from seconds to hours or days.

- ***Attacks***

Denial of Service (DoS) attacks occur when large amounts of traffic are transmitted to a particular Internet destination in an attempt to deny access to legitimate service requests. These attacks are intended to exhaust the destination's resources such as bandwidth, CPU or memory of the servers and other service-enabling devices [BGPMON]. DoS attacks can result in congestion at the intended destination and in some cases within the network that provides the destination's Internet access. In addition, these attacks can often cause congestion in a geographic region well beyond the target of the attack, for example in the networks of ISPs along the routes to the targets. Some ill-behaved applications can also mimic or have the same effect as DoS attacks. Attacks can last minutes, hours, days or even weeks.

Examples of these attacks include: SQL Slammer, a worm that spread so quickly it caused service disruptions or denial of service in large portions of the Internet as routers became overloaded and routing sessions failed [Guardian]; the DoS attacks that occurred during an Estonian government protest in 2007 [Estonia]; those that resulted from an alleged dispute between CyberBunker and Spamhaus in 2013 [Kamphuis]; and alleged ongoing attacks on financial infrastructure in North America and Europe [Atias]. An example of an application having the same impact as a DoS attack is a peer-to-peer file sharing application that consumes critical home router resources to such an extent that it may interfere with other applications.

3.2.4. Random Congestion

In addition to the events discussed above, congestion can occur because a number of users sharing a portion of a network simultaneously have high demand for a very short period of time. This random congestion is simply part of the statistical nature of traffic on the network, as illustrated previously in Figure 2. Part of the reason for this is that many applications are designed to fully utilize available resources by increasing usage up until congestion occurs, whether in the operator's network, the home or the connection between the two. This type of congestion generally has a duration from milliseconds to tenths of a second.

3.3. Location of Congestion

Congestion can occur on any link or router within the Internet. The link or router in a network path where demand is highest relative to capacity is called the *bottleneck*. Although congestion will occur on any link or router where demand exceeds capacity, it is likely that, when congestion occurs, the bottlenecks will be in relatively lower bandwidth parts of the network (access networks, for example) that connect to higher capacity parts of the network (the core ISP networks and the networks of ASPs). This follows from network design which attempts to optimize performance and cost, as capacity in access networks is generally the most expensive part of the network. Locations of potential congestion in ISP networks are:

- ***Wireless broadband access links and routers***

  Wireless access links and routers (supporting both mobile and fixed wireless broadband services) are susceptible to each of the types of congestion discussed in the previous section, particularly recurrent congestion due to busy hour demand, mass in-person events, and emergencies. Because of limited wireless spectrum, relatively high cost and complexity of adding wireless capacity, network signaling requirements, variability in bandwidth availability due to device mobility, and environmental factors, wireless access links may be the bottleneck. Congestion at these locations affects users in the geographical region served by the congested wireless access link. In addition, because wireless devices may automatically attempt to connect to any nearby access point, failures of wireless links or routers can cause the remaining links and routers to become congested, thus affecting users in a wider geographical region.

- ***Wi-Fi wireless broadband access link and routers***

  Wi-Fi wireless broadband networks are a special case of wireless broadband networks. In addition to the types of congestion faced by all wireless broadband technologies, their use of unlicensed spectrum can cause temporary reductions in capacity due to interference from adjacent Wi-Fi networks or other devices or networks operating in Wi-Fi frequencies. Congestion in a Wi-Fi network may only affect users of that network, or it may also affect users in nearby Wi-Fi networks.

- ***Wireline access links and routers***

  Wireline access links and routers are susceptible to each of the types of congestion discussed in the previous section, except those types of congestion that are caused by mobility. In particular, congestion may occur due to busy hour demand, accidents and failures, attacks, and randomness. Congestion at these locations will affect users that share the congested wireline access link or router.

- ***Core network links and routers***

  Core network links and routers are susceptible to recurrent congestion, mass distributed online events, emergencies, and accidents. However, because they have relatively high capacity and are shared by many users, they are less susceptible than

access links to attacks and to random congestion. Because traffic in the core is averaged over a greater number of users and is therefore less bursty, recurrent congestion is often minimized because it is easier to predict and plan for. When congestion occurs at these locations, however, it affects users in a wider geographical region than does congestion in access networks.

- ***Network interconnection routers***

  Network interconnection routers that connect one ISP to another ISP's network or to a large ASP's network are susceptible to recurrent congestion, mass on-line events, and attacks. Because of economies of scale, recurrent congestion is often minimized by sufficient investment in capacity. When congestion occurs at these locations, it will affect users in a wide geographical region.

There are also potential bottlenecks in users' home or office networks. These bottlenecks are not discussed in this report, as the focus here is ISP and ASP networks.

3.4. Indicators of Congestion

Network operators are continually collecting measurements on the links and routers in their networks. Due to the large volume of packets passing through a link or router, network operators will commonly aggregate measurements. For instance, rather than recording the instantaneous demand every millisecond, a network operator may calculate and record the percentile demand over a period called the *measurement interval.* The length of the measurement interval significantly affects the resulting indicator of congestion. Demand averaged over 10 second intervals will not show congestion whose duration is on time scales of milliseconds or tenths of a second, in this manner "smoothing" the resultant demand curve. Demand averaged over five-minute intervals will appear to be even smoother than demand averaged over 10-second intervals. Measurement intervals of several minutes are useful for examining congestion on time scales of minutes or longer. Common choices for measurement intervals are in the range from 5 to 15 minutes for capacity planning purposes, and many network operators examine the 95th percentile of demand over such measurement intervals.

One of the best predictors of congestion is the ratio of demand (averaged over a chosen measurement interval) to the capacity of a specific link or router, called *utilization* or *load.* When small measurement intervals are used, these measurements can be used to guide short-term congestion management. When longer measurement intervals are used, these measurements can be used to guide long-term congestion management.

Figure 5 shows an example of the percentage of link utilization on an access link interface during a one-hour time period using different measurement intervals. In addition to measuring demand, network operators commonly measure the average time from the arrival to the departure of a packet at a link or router (delay), and the proportion of packets that are not transmitted (packet loss).

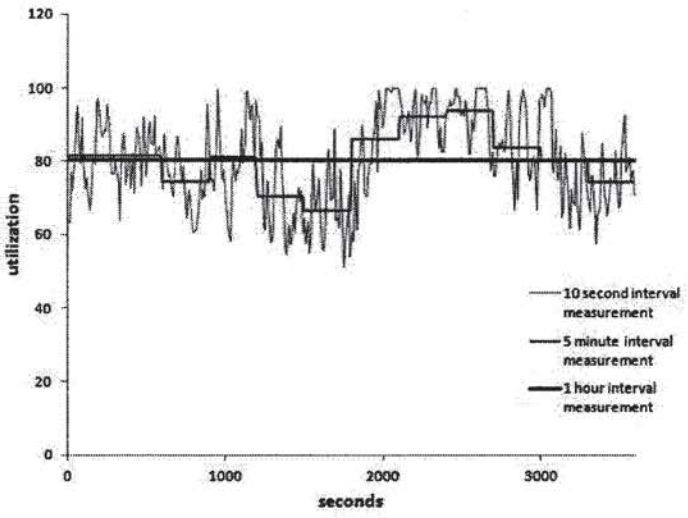

Figure 5. Link Utilization for Various Measurement Intervals

Congestion can also be measured by the effect that it has on a user's application. These are called *Quality of Service (QoS)* metrics. The most common are:

- *End-to-end packet delay*: The time from the transmission of a packet at the source to its reception at the destination. This includes delay due to (a) the time for a signal to propagate along the links, (b) the time for transmission of a packet at each router, and (c) queuing time due to congestion.
- *Delay jitter* (or simply "jitter"): The variation in end-to-end delay between packets.
- *End-to-end packet loss*: The proportion of packets that are transmitted by a source that do not arrive at the destination.
- *End-to-end throughput*: The average number of bits per second that are received at the destination.

3.5. Impact of Congestion on and by Applications

The satisfaction of a user with the performance of an application is called *Quality of Experience (QoE)*. When increased end-to-end packet delay, delay jitter, or end-to-end packet loss cause a degradation in QoE, it is generally noticed by the user in a variety of ways. Examples include:

- Increased response time of all or specific Internet applications.
- Webpages or parts of webpages (images, for example) take an increased time to load.
- Streaming audio or video suffers from decreased sound or picture quality, or is interrupted.

- Real-time audio (such as voice calls) or real-time video (such as video chat) suffers from decreased sound or picture quality or from unacceptable delays between speaking and hearing, or is interrupted.
- In multiplayer games, players may notice an increased delay between actions taken on the controller or home device and the results of these actions on the screen.
- Increased file transfer times.

Congestion may or may not be noticeable to users, depending on the application type, the application design, the severity of congestion as measured by QoS metrics, and the duration of congestion. The characteristics of an application and its design determine when and how QoE is degraded by congestion. Applications can be roughly classified by their sensitivity to QoS metrics and to the duration of congestion:

- ***Delay-intolerant applications:***

  Applications that are highly interactive are likely to have a QoE that is very sensitive to end-to-end delay. For example, the QoE of voice calls, video chat, and many multi-player games suffers when the end-to-end delay exceeds a few tenths of a second. Consequently, delay-intolerant applications will usually not request that dropped packets be retransmitted from the sender and will usually throw away packets that do not arrive within a certain time interval. Even brief occurrences of congestion, on the order of a few tenths of a second, can cause noticeable degradation in the QoE of these applications.

- ***Jitter-intolerant applications:***

  Applications that support synchronous communication are likely to have a QoE that is sensitive to delay jitter on the order of a few seconds or less. Since jitter is caused by variations in delay, all delay-intolerant applications are also jitter-intolerant. In addition, applications that stream audio and video are often jitter-intolerant. Consequently, jitter-intolerant applications will buffer received packets to equalize their end-to-end delay and thereby reduce their delay jitter. Real-time communications applications will only buffer packets for at most a few tenths of a second. In contrast, streaming applications often buffer packets for a few seconds. Time periods in which instantaneous demand exceeds capacity, that last longer than the buffering capability of a jitter-intolerant application, can cause noticeable degradation in its QoE.

- ***Minimum throughput applications:***

  Some applications are designed with the assumption that they will experience at least a certain minimum throughput. For example, the QoE of voice and video often degrades significantly if the throughput falls below the designed threshold. If the content is encoded in multiple formats, the application may respond to congestion by changing to a format with a lower throughput threshold, which reduces the QoE but is often preferable to not receiving the content at all. Minimum throughput applications can

experience noticeable degradation in QoE when instantaneous demand exceeds capacity during time periods on the order of tens of seconds.

- ***Loss-intolerant applications:***

  Applications have different tolerances for end-to-end packet loss. Applications that expect data to arrive accurately and without error do not tolerate packet loss. Streamed video is an example of this, as users find missing pixels, frozen frames, and other manifestations of lost packets to provide a very unpleasant QoE. Other applications can tolerate a small amount of end-to-end packet loss, for example some audio and video applications. Applications that are loss-intolerant but delay jitter-tolerant, such as text messaging, text chat, email, and web browsing, will request that dropped packets be retransmitted to ensure that every packet is eventually received at the destination, even if some packets require multiple attempts. Degradation in the QoE of these applications will typically occur only if the duration of congestion is seconds to minutes. Congestion of any duration may delay file transfer times for bulk data transfer applications (such as software updates or peer-to-peer file sharing), with delays in proportion to the duration of congestion.

In summary, occurrences of congestion for a few tenths of a second may degrade delay-intolerant applications, occurrences of congestion for a few seconds may degrade jitter-intolerant applications, and occurrences of congestion for tens of seconds may degrade minimum-throughput applications. If the duration of congestion is on the order of minutes to hours, then it negatively impacts most applications. An operator's network design as well as congestion management practices affect whether and how congestion impacts various applications.

Network design can significantly affect the QoE of various applications. For example, the size of buffers in routers can affect the delay and loss of incoming traffic to each queue (see Section 5.7 on packet scheduling). The use of larger queues can decrease packet loss but increase delay and jitter.

The behavior of applications can either decrease or increase congestion. Some applications respond to congestion by decreasing their sending rates, thereby decreasing congestion. Some delay-tolerant applications may not decrease their sending rates but may use protocols that allow traffic transmissions to be scheduled in a manner that decreases congestion. There are other applications, however, that do not react to congestion or that react in a manner that increases traffic, e.g. by requesting retransmission of all delayed or dropped packets, without lowering the rate of transmission.

4. Classification of Congestion Management Techniques and Practices

This section of the report provides a classification of congestion management techniques and practices and delineates the scope of the report. Section 4.1 distinguishes between congestion management techniques and practices, or in other words the application of those techniques to effectuate a particular outcome. Section 4.2 classifies congestion management techniques by the time scale on which they operate. Section 4.3 briefly outlines the parties or organizations that may implement congestion management practices. Section 4.4 discusses when congestion management practices are based on user- or application- based criteria. Section 4.5 delineates the scope of the remainder of the report.

4.1. Techniques versus Practices

It will be helpful in this report to distinguish between congestion management techniques and specific implementations of those techniques, which are referred to as "practices", or in other words the application of those techniques to effectuate a particular outcome. This report uses the term *congestion management technique* to refer to a specific congestion management function that determines whether Internet traffic is transmitted or the rate at which traffic is transmitted, or that enables such functionality in other techniques. The techniques considered in this report include packet scheduling, packet dropping, routing, rate control, caching, resource reservation, and admission control, which are discussed in Section 5.

This report uses the term *congestion management practice* to refer to the use by a party or organization:

- of a collection of traffic management techniques,
- targeted at particular users and/or applications,
- upon the trigger of some event.

The parties or organizations that engage in or implement congestion management practices include Internet Service Providers (ISPs), Application Service Providers (ASPs), operating systems developers, customer premises equipment manufacturers, and consumers and enterprises. For example, an ISP may combine a set of congestion management techniques with the goal of reducing congestion for all users and applications, or with the goal of reducing congestion only for a subset of users or applications. In the latter case, congestion management may be used to differentiate products. An ASP that operates a network, for example, may combine a set of congestion management techniques with the goal of reducing congestion for all of its users and applications, or with the goal of reducing congestion only for a subset of its users or applications.

The Internet is based on the concept of a layered architecture, where each layer provides certain functionalities [RFC 1122]. A *layer* is an abstraction that hides the implementation details of a particular set of functionality. Congestion management can be applied in any of the layers. (A definition of each layer can be found in the Glossary.) Packet scheduling is

commonly implemented in the lowest three layers: the *physical*, *link*, and *network* layers. Packet dropping and routing are commonly implemented in the network layer. Rate control and admission control are commonly implemented in the highest two layers: the *transport* and *application* layers. Caching is commonly implemented in the application layer. Resource reservation may be implemented at any layer, but is often controlled by decisions at the transport and application layers. Techniques that use deep packet inspection (DPI) are usually operating at multiple layers.

Section 5 of this report presents a survey of congestion management techniques to illustrate the range of techniques used by network operators. Section 6 presents a survey of congestion management practices of network operators.

4.2. Time Scales

Since congestion occurs on time scales from milliseconds to hours, as discussed in Section 3.2, congestion management techniques are also designed to work on a number of different time scales.

- ***Months to Years – Capacity Planning; Internet Subscription Plans***

  Capacity planning and augmentation occurs on a time scale of months to years. ISPs face rapidly growing demand for capacity. During the last year, average Internet demand per customer during the busiest hours in North America grew at an annual rate of 39% on fixed access lines and 25% on mobile access networks according to one estimate [Sandvine2013]. During the next five years, average Internet demand during the busiest hours in North America is expected to grow at an annual rate of 23% according to one estimate [CiscoVNI] and mobile Internet traffic at an annual rate of 40% according to another [Sandvine2013]. Network operators typically estimate demand months to years in advance, and use such demand estimates combined with the cost of capacity to plan a schedule for capacity upgrades. Since it may take months to implement a capacity upgrade, the time scale for such congestion management is months to years. Thus, although capacity planning can greatly affect how much congestion occurs on multiple time scales, it cannot react to congestion as it occurs.

  Internet subscription plans also affect congestion on a time scale of months to years. Internet subscription plans commonly include limits on downstream and upstream transmission rates. Since demand usually increases with transmission rates, the number of subscribers to a particular plan can affect demand and thus congestion. In addition, some plans include limits on the maximum number of bytes transmitted per month (commonly called *usage caps*) or charges for usage above some threshold. These limits can be viewed as long-term congestion management practices. Limits on transmission rates affect congestion in much the same way as do capacity decisions. Usage caps and usage charges may influence the amount of traffic users transmit over the course of a billing cycle. Although such limits can affect how often or the degree to

which congestion occurs on multiple time scales, they cannot react to congestion as it occurs.

- ***Seconds to Minutes – Reservation; Prioritization; Rate Control; Routing***

Reservation, prioritization, and admission control techniques and practices can be used to react to congestion on a time scale of seconds to minutes, as well as to differentiate or guarantee performance on shorter time scales. Some networks allow resources such as bandwidth to be reserved or allow some packets to be prioritized over others. Prioritization techniques are usually implemented through packet scheduling in the link or network layers. Resource reservation may be implemented in any layer, but is often controlled by decisions at the transport or application layers. Typically the decisions that guide use of such practices are made on a time scale of seconds to minutes.

Rate control techniques such as those implemented in the Transmission Control Protocol (TCP) are designed to react to congestion on a time scale of a few tenths of a second or longer. Congestion management protocols residing at the transport or application layers are often used to limit the number of packets per second that a source application transmits. These congestion management techniques use information about end-to-end packet delays and losses, and thus dynamically update their limits on the time scale of an end-to-end delay (typically on the order of a few tenths of a second).

Although routing is typically based solely on the destination IP address, routes are sometimes adjusted in an attempt to minimize congestion. When these adjustments are made, typically it is on a time scale of minutes or longer. Packets generally progress through many routers before arriving at their destination. The path is determined by routers that exchange information concerning possible routes and congestion on these routes. In addition, sometimes content is available at multiple locations, and Content Delivery Networks can be used to balance the load and reduce congestion. This computation of routes is accomplished using network layer protocols. Routes may be updated as often as every few tenths of a second.

- ***Fractions of a Second – Packet Scheduling; Packet Dropping***

Packet scheduling techniques can be used to react to congestion on a very fast time scale. Packet scheduling techniques at the *data link layer* determine when to transmit each packet. When there is a queue of packets waiting for transmission, packet scheduling techniques also choose which packet to transmit next. This decision takes place each time a packet is transmitted, which can be roughly 100 to 1 million times per second, depending on the transmission rate and the packet size. The decision is guided by a simple algorithm that requires little computation. When a queue is full or nearly full, packets might be dropped. This decision takes places each time a packet arrives in a queue, which might also be roughly 100 to 1 million times per second.

4.3. Parties that May Engage in Congestion Management Practices

The parties or organizations that may engage in congestion management practices are:

- **Internet Service Providers**

  Each Internet Service Provider (ISP) implements a set of congestion management techniques at each router within its network. Thus, all applications communicating through the Internet indirectly rely on congestion management techniques implemented within networks of the ISPs through which traffic passes. Congestion management techniques within ISPs' networks includes routing (Section 5.5), packet dropping (Section 5.6), and packet scheduling (Section 5.7). Optionally, an ISP may also support admission control and resource reservation (Section 5.2). Certain networks, notably cellular data networks, also implement rate control (Section 5.4). Some ISPs also use deep packet inspection to classify traffic (Section 5.1).

- **Application Service Providers and Application Designers**

  Applications can, at times, implement congestion management techniques at each endpoint of the communication. At the user's endpoint, congestion management may be implemented within an application that a user runs on a device. At remote endpoints, congestion management may be similarly implemented in the communicating application at another location, within an application server, or within an ASP's own network. Congestion management is frequently implemented by applications that are moderately to highly interactive, including video streaming, voice over IP (VoIP), video conferencing, and gaming. Congestion management techniques within applications may include admission control and resource reservation (Section 5.2), caching (Section 5.3), and rate control (Section 5.4).

- **Operating Systems Developers**

  Each device's operating system (e.g. Windows, MacOS, Linux, Android, iOS) implements a set of congestion management techniques at each endpoint of the communication. Since all applications communicate by relying on the network functionality built into the operating system, all applications indirectly rely on congestion management techniques implemented within the operating system. Congestion management techniques within operating systems may include protocol support for allowing applications to request admission control and resource reservation (Section 5.2), application interfaces to TCP rate control capabilities (Section 5.4), and allowing applications to access and control protocol header information that can impact packet dropping (Section 5.6) and packet scheduling (Section 5.7).

- **Customer Premises Equipment Manufacturers**

  Customer premises equipment (CPE) consists of user end devices (including computers, smartphones, tablets, Internet-connected printers, Internet-connected digital video recorders, and Internet-connected game systems) and user networking devices (including cable modems, DSL modems, home routers, and home gateways). Each piece

of CPE implements a set of congestion management techniques. Thus, all applications communicating from one device to another or to a server through the Internet indirectly rely on congestion management techniques implemented within CPE. Congestion management techniques implemented within CPE as part of the device driver includes packet scheduling (Section 5.7).

- **Consumers and Enterprises**

Both residential consumers and enterprises may implement congestion management techniques in their networks. For both consumers and enterprises, such techniques may be activated and configured within applications, operating systems, and CPE. Large enterprises may also implement congestion management techniques in a manner similar to ISPs.

4.4. Which Traffic is Subject to Congestion Management

One of the key design questions about any congestion management practice relates to the subset of network traffic with which the practice is concerned. Network operators target all traffic on their networks with some practices, whereas with other practices they target only the traffic of specific users, a subset of types of applications, all instances or specific instances of applications, or specific components of such applications.

User-based congestion management is applied to all traffic associated with a particular user or user group. Some ISPs or ASPs may define user groups based on:

- the service plan to which users are subscribed (e.g., all users subscribed to an ISP's basic broadband Internet access plan, a cellular provider's unlimited data plan, or an ASP's premium product);
- the volume of data that users send or receive over a specified period of time or under specific network conditions (e.g., all users who consume 300 GB in a month, the top 5% of users by data consumption during a busy period on the network, or users who are in the process of transmitting the first 20MB of a file); or
- the location of users (e.g., all users in a particular geographic area experiencing an emergency).

User-based congestion management does not require network operators to examine the content of network traffic as the decision to apply user-based management is agnostic to that content – it depends only on which users are generating traffic, not what they are generating.

Application-based congestion management is applied to all traffic associated with particular uses of the network. That is, congestion management is application-based if network operators select traffic to be managed because it:

- has a particular source or destination (e.g., http://www.example.com);
- is generated by a particular application (e.g., a BitTorrent client);
- is generated by an application that belongs to a particular class of applications (e.g., video chat applications that include Skype, Google Talk, WebEx, and FaceTime);
- uses a particular application- or transport- layer protocol (e.g., Session Initiation Protocol, User Datagram Protocol, or Hypertext Transfer Protocol); or
- is classified for special treatment by the user, application, or application provider (e.g. traffic identified by the user's application as delay-intolerant, or traffic identified by the application provider as jitter-intolerant).

Application-based congestion management depends on the network operator's ability to identify the traffic associated with particular uses of the network. Techniques used by network operators to identify and select traffic subject to an application-based congestion management practice might be based on packet payloads (using deep packet inspection or other content-aware network equipment), network or transport layer headers (e.g., port numbers or priority markings), heuristics (the size, sequencing, and/or timing of packets), or a combination of these characteristics.

Congestion management may also be both user- and application- based. For example, a network operator could choose to rate-limit video streaming for all users who consume 300 GB in a month or for all users in a congested cell in a cellular network.

User- and application- based congestion management may be based in part on economic and legal agreements between network operators or between users and network operators. *Service Level Agreements* (SLAs) between network operators delineate contractual aspects of the service, often including the upstream and downstream bit rates at the boundary between the operators' networks, the maximum delay across an operator's network, sometimes the maximum proportion of packets that may be dropped or other QoS metrics, and sometimes specifications of payments.

In contrast to both user-based and application-based management, some practices apply to all traffic regardless of user group or application. For example, a network operator might program its routers with dynamically adjusting buffers to accommodate rapid changes in network load (Section 5.6). This practice applies to all traffic on the network and can help to mitigate the impact of congestion regardless of which users' traffic is flowing through the routers or which applications are in use. Similarly, network operators might make decisions about traffic scheduling, queuing, or routing based on factors unrelated to users or applications – how quickly packets have arrived at a particular point in the network, or which ones arrived first, for example. These kinds of practices are user- and application-agnostic.

4.5. Scope of the Remainder of the Report

The remainder of this report focuses on user- or application- based real-time network management of Internet services. This is an important set of congestion management techniques and practices. It is also a set that has generated public policy discussions. In order to be considered in the remainder of this report, a congestion management technique or practice must:

(1) Be implemented or potentially implemented by a network operator. As discussed in Section 4.3, other parties or organizations that may implement congestion management practices – such as applications, operating systems, and device drivers – are important; however they are outside the scope of this report.

(2) Act on a time scale of minutes or less. As discussed in Section 4.2, congestion management techniques and practices that operate on a time scale greater than minutes (e.g. capacity upgrades, limits on downstream and upstream transmission rates, usage caps and usage charges) may influence the amount of traffic users transmit over the course of a billing cycle; however they are outside the scope of this report.

(3) Be used for purposes of congestion management. Similar techniques may also be used for other purposes, including security; however use for such non-congestion management purposes is outside the scope of this report.

(4) Be user- or application- based. Congestion management techniques and practices whose goal is to reduce congestion for a chosen set of users or applications will be considered. As discussed in Section 4.4, there are many congestion management techniques whose goal is to reduce congestion for all users and all applications; however they are outside the scope of this report.

This report focuses on congestion management techniques and practices that pass all four of these tests. However, the BITAG notes that congestion management techniques and practices that do not pass all four of these tests are important, and it may consider them in future reports.

5. Congestion Management Techniques

As noted in Section 4.1, this report uses the term *congestion management technique* to refer to a specific congestion management function that determines whether Internet traffic is transmitted or the rate at which traffic is transmitted, or that enables such functionality in other techniques. Many congestion management techniques at or above the network layer are standardized by the Internet Engineering Task Force (IETF), and many at the physical and data link layers are standardized by the Institute of Electrical and Electronics Engineers (IEEE) or by industry consortia (e.g. CableLabs, 3GPP). Standardization can allow for interoperability and for consistent functionality in network devices and equipment. Other congestion management techniques have not been standardized or are propriety.

This section focuses on congestion management techniques and for each technique explains who may apply the technique, the duration and location of congestion the technique addresses, and the intended impact upon applications. Section 5.1 illustrates how either a user or a network operator may classify a packet. Section 5.2 discusses congestion management techniques designed to support applications that require a minimum amount of network resources in order to function at the desired performance level. Section 5.3 discusses congestion management techniques that temporarily store (or *cache*) popular content in multiple locations. Section 5.4 discusses congestion management techniques that control the average rate at which a source transmits traffic into the Internet. Section 5.5 considers routing and traffic engineering. Section 5.6 discusses how a router may decide when to drop a packet and/or *mark* it to indicate congestion. Section 5.7 examines how a router places packets into queues and the order in which it transmits packets. Finally, Section 5.8 discusses how these techniques may be combined to offer a collection of capabilities in various QoS architectures and in various access network architectures.

Each of these sections focuses on congestion management techniques that may be user- or application- based. Congestion management techniques that are agnostic to both user and application are very commonly implemented but are not generally discussed here. Section 6 illustrates congestion management practices that use many of these techniques.

5.1. Packet Classification

In order for a congestion management technique to be based on a particular user or a particular application type, application, or application component, packets must be classified using criteria that identifies the particular user or application traffic. This section illustrates how users, ASPs, and ISPs can classify packets.

Most packets in the Internet are transmitted by network operators using *best-effort service*, in which routers transmit packets in the order in which they are received, and without regard to the source, the application that generated them, or the resulting QoE (as discussed in Section 3.5). Hence best-effort service requires only knowledge of the destination of a packet, not of the source, the application, or the desired QoS characteristics.

In contrast, user- or application- based congestion management may require knowledge not only of the destination but also of the source, user behavior, the type of application, application, application component, the user's or application's desired QoS characteristics or QoE, agreements between network operators, or agreements between users and network operators (as discussed in Section 4.4). Each user, ASP, and ISP may classify packets on the basis of such information. A *classifier* is an entity that selects packets based on the content of packet headers or other attributes according to defined rules [RFC2475]. Classification of packets may be performed using attributes from any of the following layers: application, transport, network, or data link. This report uses the term *flow* to

indicate a group of packets that share a common set of properties [RFC5472]. One purpose of classification is to allow a network operator to apply congestion management techniques based in part on a packet's classification.

A packet's classification is often used to decide how to apply congestion management techniques discussed in the remainder of this section. Classification permits the network operator to make choices, for example ensuring that applications receive the desired treatment, to force "aggressive" applications (e.g., applications that attempt to use all available resources) to make way for less aggressive applications, or to ensure that bandwidth contracted for in an SLA is available for the intended purpose.

Markings that indicate the classification of a packet are often placed in specific bits of a packet header to allow simplified subsequent classification based only on these specific bits. The classification markings can allow a user, ASP, or ISP to indicate a desired packet treatment. For example, a source can use a classification marking to request that the packet be treated in a manner consistent with expectations for Voice-over-IP traffic [RFC 2474, 2475, 4594], as discussed below in Section 5.8.2. Alternatively, the classification marking can be used to indicate packets to or from a specific host or network, or traffic related to a given application, as discussed below in Sections 5.2 and 5.5. Classification markings can also be used to uniquely identify a flow, so that priority can be given on the basis of the throughput experienced by a flow.

A classification marking can be placed in the packet header corresponding to the protocol that generated it. The classification marking may indicate a unique identifier of the flow. For example, the Multiprotocol Label Switching (MPLS) protocol (discussed below in Section 5.5) includes fields in its packet header that can be used to classify packets and to assign the route (or portion of a route) taken by those packets. Alternatively, the classification marking may indicate the desired treatment of the packet. For example, the DiffServ architecture (discussed below in Section 5.8.2) includes a *codepoint* in the IP packet header that can be used to classify packets of *differentiated services* (e.g., network control, VoIP, video streaming, best effort traffic) in order to provide desired routing, scheduling, and dropping treatment. Data link layer protocols often include similar classifications, e.g. the Ethernet protocol includes a field in its packet header that can be used to classify packets for similar purposes [802.1Q][802.1p].

Each user, ASP, and ISP may classify, mark, and re-mark packets according to its architecture, objectives, and agreements. Thus a packet's classification marking may be modified as it progresses from source to destination. Most commonly, marking and re-marking may occur at user-operator and operator-operator boundaries, according to the agreement or lack thereof between the parties. For instance, a network operator may classify a packet upon ingress to its network by inspecting the MPLS fields, the DiffServ codepoint, the Ethernet fields, or another data link layer codepoint. It may also classify packets based on the transport protocol, source IP address, source port, destination IP address, or destination port. If a network operator does not support packet classification or does not have an agreement with a user, ASP, or ISP to honor a packet's classification